Davide Belli
11887532
davidebelli95@gmail.com

Homework Assignment 1
Machine Learning 2, 17/18

2018-04-15

In this Homework, I discussed possible interpretations with Gabriele Cesa.

1 Problem 1

1.1 Question 1

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) = \\ &= H(Y) - H(Y|X) = \\ &= \mathrm{KL}\big(p(x,y)||p(x)p(y)\big) \end{aligned}$$

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) = \\ &= H(Y|Z) - H(Y|X,Z) = \\ &= \mathbb{E}_Z\Big[\mathrm{KL}\big(p(x,y|z)||p(x|z)p(y|z)\big)\Big] \end{aligned}$$

Mutual Information measures the amount of Information shared between two variables. In other words, how much information about X do I already know if Y is given? The same is valid viceversa, since Mutual Information is symmetric. Conditional Mutual Information measures Mutual Information between two variables if another variable value is given. More formally, CMI represent the expected value of the mutual information of X,Y random variables given Z. Notice that, if both X and Y are independent from Z, CMI equals MI.

1.2 Question 2

$$\begin{aligned} I(X,Y) &= \mathrm{KL}\big(p(x,y)||p(x)p(y)\big) \\ &= \sum_{x\in X}\sum_{y\in Y} p(x,y)\Big(\log p(x,y) - \log p(x)p(y)\Big) \\ &\approx -0.0187 + 0.0200 + 0.0200 - 0.0180 \\ &= 0.0033 > 0 \end{aligned}$$

To find this result, I used computations described in Table 1. If the mutual information between two variables is greater than zero, it means that, given the value of one of the two variables, we have additional information about the possible values of the other one. More clearly, the entropy of a variable given the value of the other one is less than its entropy without knowing the value of the second variable, as expressed in: $I(X;Y) = H(X) - H(X|Y) > 0 \implies H(X) > H(X|Y)$. The same applies for Y given X.

x	y	$p(x,y)$	$p(x)$	$p(y)$	$\log p(x,y) - \log p(x)p(y)$	$p(x,y)\Big(\log p(x,y) - \log p(x)p(y)\Big)$
0	0	0.336	0.6	0.592	-0.0556	-0.0187
0	1	0.264	0.6	0.408	0.0755	0.0200
1	0	0.256	0.4	0.592	0.0780	0.0200
1	1	0.144	0.4	0.408	-0.1250	-0.0180

Table 1

1.3 Question 3

$$\begin{aligned} I(X,Y|Z) &= \sum_{z\in Z} p(z)\mathrm{KL}\big(p(x,y|z)||p(x|z)p(y|z)\big) \\ &= \sum_{z\in Z} p(z)\sum_{x\in X}\sum_{y\in Y} p(x,y|z)\Big(\log p(x,y|z) - \log p(x|z)p(y|z)\Big) \end{aligned}$$

$$= 0 + 0 = 0$$

To find this result, I used computations in Table 2, where the values in the latest columns are:
$V_1 = \log p(x, y|z) - \log p(x|z)p(y|z)$
$V_2 = p(x, y|z)\Big(\log p(x, y|z) - \log p(x|z)p(y|z)\Big)$
$V_3 = \text{KL}\big(p(x, y|z)||p(x|z)p(y|z)\big)$
$V_4 = p(z)\text{KL}\big(p(x, y|z)||p(x|z)p(y|z)\big)$

If the conditional mutualinformation $I(X, Y|Z)$ is zero it means that, given the value of the variable Z, Y does not carry additional information about X (that is not already included in Z), and X does not carry any additional information about Y. Precisely: $I(X; Y|Z) = H(X|Z) - H(X|Y, Z) = 0 \implies H(X|Z) = H(X|Y, Z)$. The same applies for Y.

x	y	z	$p(x, y\|z)$	$p(x\|z)$	$p(y\|z)$	V_1	V_2	V_3	V_4
0	0	0	0.4	0.5	0.8	0	0	0	0
0	1	0	0.1	0.5	0.2	0	0		
1	0	0	0.4	0.5	0.8	0	0		
1	1	0	0.1	0.5	0.2	0	0		
0	0	1	0.277	0.692	0.4	0.00072	0.00020	0	0
0	1	1	0.415	0.692	0.6	-0.00048	-0.00020		
1	0	1	0.123	0.308	0.4	-0.00162	-0.00020		
1	1	1	0.185	0.308	0.6	0.00108	0.00020		

Table 2

1.4 Question 4

We show in Table 3 that, for every possible value combination of the random variables, the following holds:

$$p(x)p(z|x)p(y|z) = p(y, z|x)p(x) = \\ = p(x, y, z)$$

x	y	z	$p(x)$	$p(z\|x)$	$p(y\|z)$	$p(x)p(z\|x)p(y\|z)$
0	0	0	0.6	0.4	0.8	0.192
0	1	0	0.6	0.4	0.2	0.048
1	0	0	0.4	0.6	0.8	0.192
1	1	0	0.4	0.6	0.2	0.048
0	0	1	0.6	0.6	0.4	0.144
0	1	1	0.6	0.6	0.6	0.216
1	0	1	0.4	0.4	0.4	0.064
1	1	1	0.4	0.4	0.6	0.096

Table 3

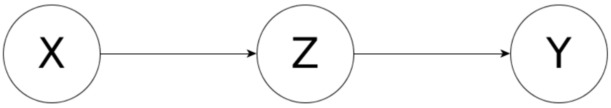

Figure 1: DAG for $p(x)p(z|x)p(y|z)$

2 Problem 2

To shorten the set of independence rules for every cluster, we will use the following notation to say that X, Y are independent when conditioned from any other variable (including the empty set of variable). In our case,

there are only three variable, thus we want to say that that X and Y are independent and independent given Z.
$X \perp\!\!\!\perp Y|*$
The same holds for not independence.
Also, in order to avoid plotting more than 20 graphs, some cases similar cases (e.g. same graph with an arrow in the opposite direction) are not plotted by explained in words.

Finally, notice that cluster sets {2,3,4}, {5,6,7}, {8,9,10} can be clustered together in three new supersets if we consider the possibility to switch node names on the graph. In other words, those sets of graphs represent the same dependence conditions but applied to different variables.
Cluster 1 (Fig. 2)
$X \perp\!\!\!\perp Y| *$
$X \perp\!\!\!\perp Z| *$
$Y \perp\!\!\!\perp Z| *$

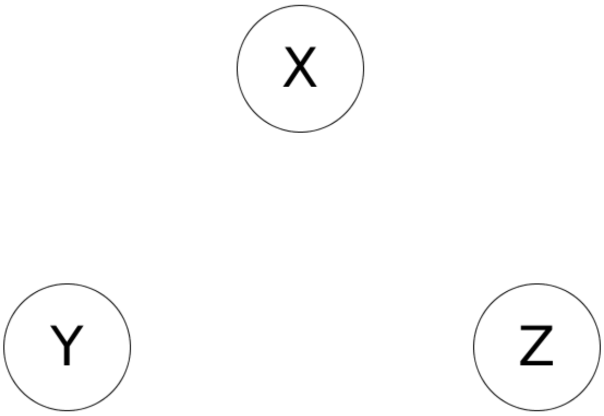

Figure 2

Cluster 2 (Fig. 3)
$X \not\perp\!\!\!\perp Y| *$
$X \perp\!\!\!\perp Z| *$
$Y \perp\!\!\!\perp Z| *$
This holds also with the arrow from X to Y in the opposite direction.

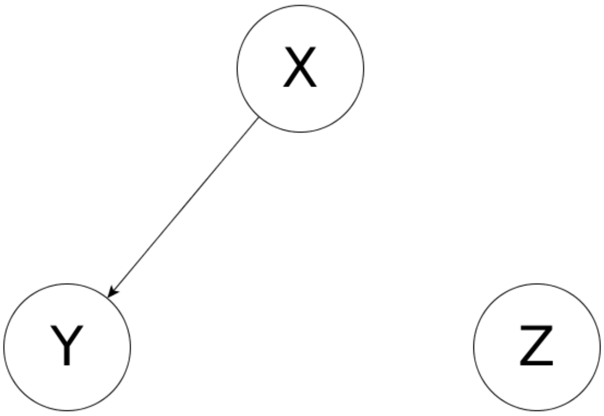

Figure 3

Cluster 3 (Fig. 4)
$X \perp\!\!\!\perp Y| *$
$X \perp\!\!\!\perp Z| *$
$Y \not\perp\!\!\!\perp Z| *$
This holds also with the arrow from Y to Z in the opposite direction.

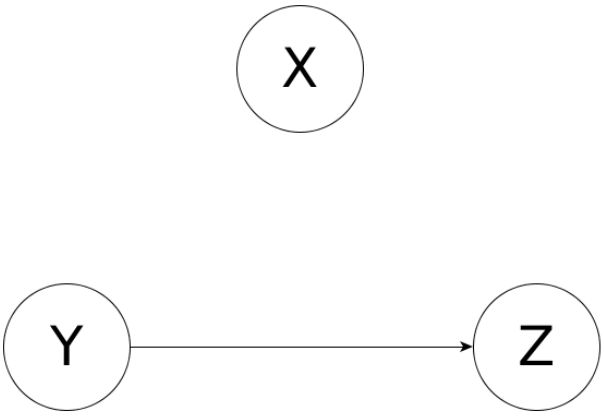

Figure 4

Cluster 4 (Fig. 5)
$X \perp\!\!\!\perp Y | *$
$X \not\perp\!\!\!\perp Z | *$
$Y \perp\!\!\!\perp Z | *$
This holds also with the arrow from Z to X in the opposite direction.

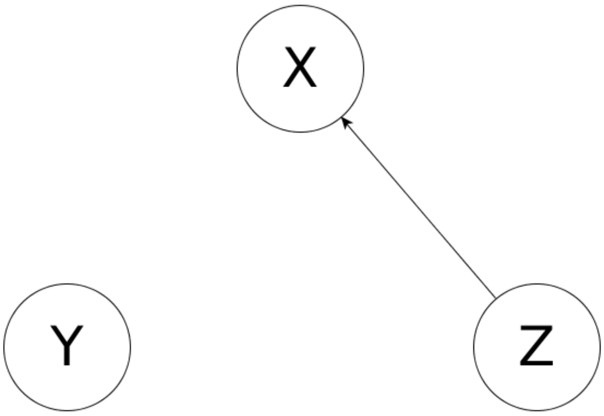

Figure 5

Cluster 5 (Fig. 6)
$X \not\perp\!\!\!\perp Y | *$
$X \not\perp\!\!\!\perp Z | *$
$Y \perp\!\!\!\perp Z | X$
$Y \not\perp\!\!\!\perp Z$
This holds also with both the arrows in the chain case in opposite direction.

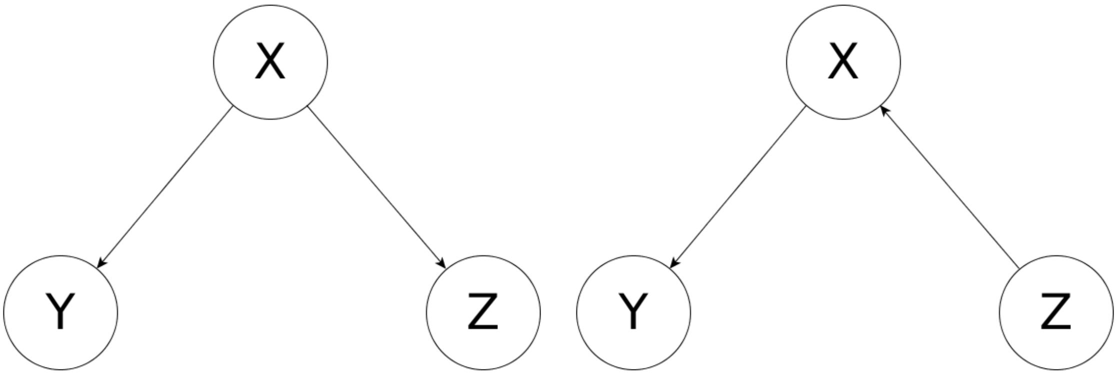

Figure 6

Cluster 6 (Fig. 7)
$X \not\!\perp\!\!\!\perp Y| *$
$X \perp\!\!\!\perp Z|Y$
$Y \not\!\perp\!\!\!\perp Z| *$
$X \not\!\perp\!\!\!\perp Z$
This holds also with both the arrows in the chain case in opposite direction.

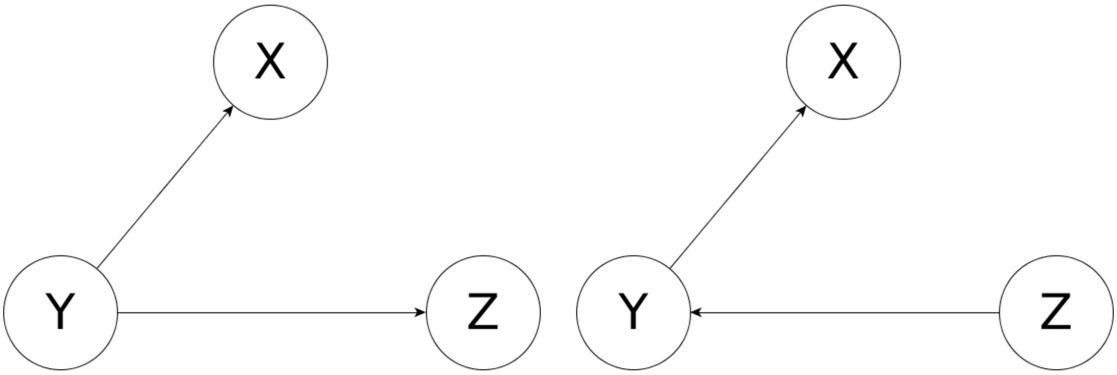

Figure 7

Cluster 7 (Fig. 8)
$X \perp\!\!\!\perp Y|Z$
$X \not\!\perp\!\!\!\perp Z| *$
$Y \not\!\perp\!\!\!\perp Z| *$
$X \not\!\perp\!\!\!\perp Y$
This holds also with both the arrows in the chain case in opposite direction.

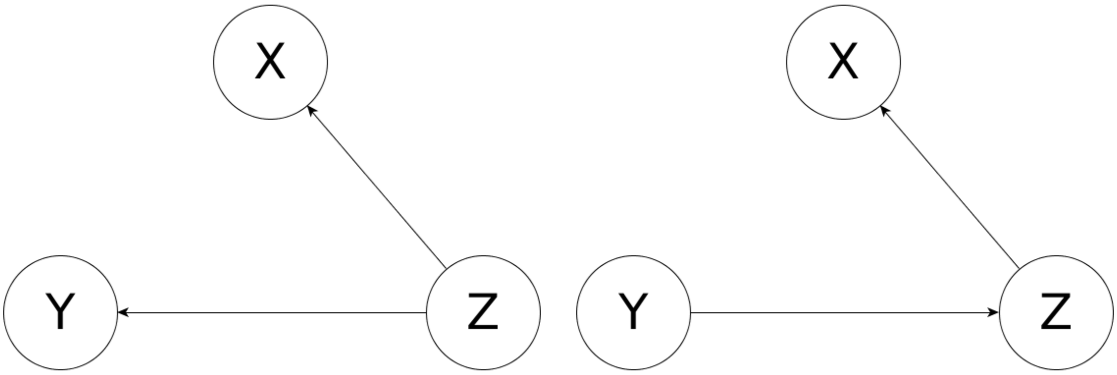

Figure 8

Cluster 8 (Fig. 9)
$X \perp\!\!\!\perp Y|*$
$X \perp\!\!\!\perp Z|*$
$Y \not\!\perp\!\!\!\perp Z|X$
$Y \perp\!\!\!\perp Z$

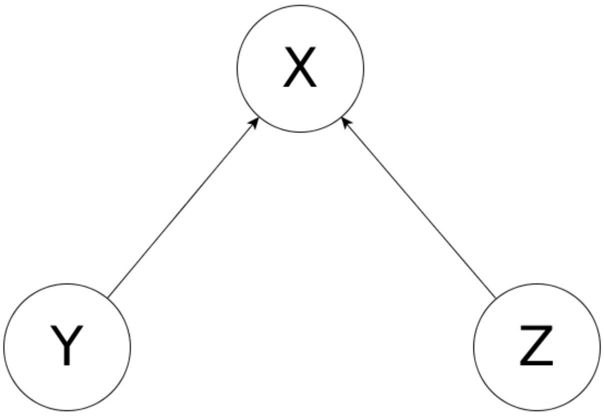

Figure 9

Cluster 9 (Fig. 10)
$X \not\!\perp\!\!\!\perp Y|*$
$X \not\!\perp\!\!\!\perp Z|Y$
$Y \not\!\perp\!\!\!\perp Z|*$
$X \perp\!\!\!\perp Z$

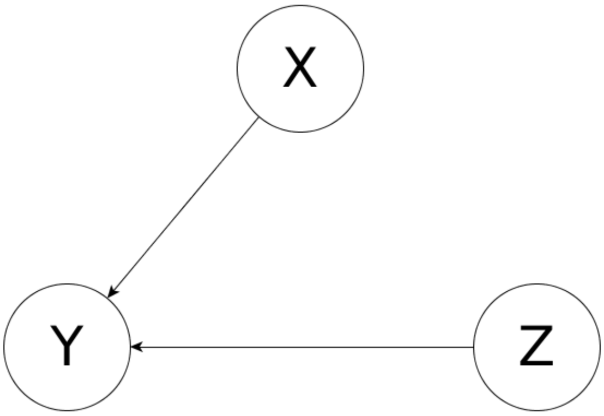

Figure 10

Cluster 10 (Fig. 11)
$X \not\!\perp\!\!\!\perp Y|Z$
$X \not\!\perp\!\!\!\perp Z|*$
$Y \not\!\perp\!\!\!\perp Z|*$
$X \perp\!\!\!\perp Y$

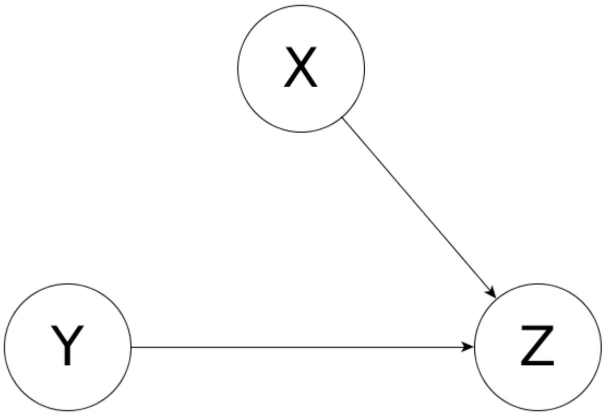

Figure 11

Cluster 11 (Fig. 12)
For every DAG with three edges between the three vertices: $X \not\perp\!\!\!\perp Y | *$
$X \not\perp\!\!\!\perp Z | *$
$Y \not\perp\!\!\!\perp Z | *$

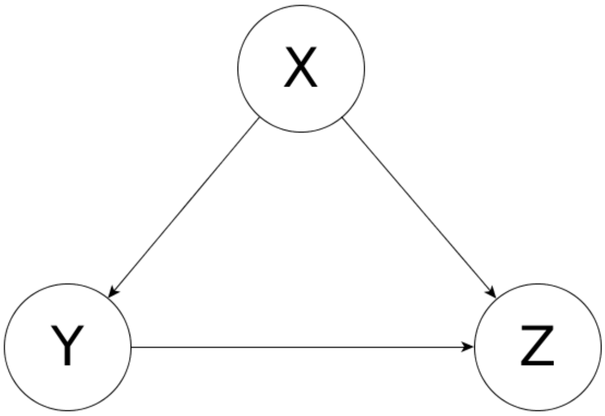

Figure 12

3 Problem 3

3.1 Question 1

Notice that for this Question we will be using solutions of Problem 3, Question 2 for the entropy of p.

$$
\begin{aligned}
\mathrm{KL}(p||q) &= -\int p(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} dx = \\
&= -\int p(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} dx = \\
&= \int p(\boldsymbol{x}) \log p(\boldsymbol{x}) dx - \int p(\boldsymbol{x}) \log q(\boldsymbol{x}) dx = \\
&= -H(p) - \int p(\boldsymbol{x}) \log q(\boldsymbol{x}) dx = \\
&= -\frac{D}{2} \log 2\pi - \frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{D}{2} + \frac{D}{2} \log 2\pi + \frac{1}{2} \log |\boldsymbol{L}| + \frac{1}{2} \int p(\boldsymbol{x}) (\boldsymbol{x} - \boldsymbol{m})^T \boldsymbol{L}^{-1} (\boldsymbol{x} - \boldsymbol{m}) dx = \\
&= \frac{1}{2} \log \frac{|\boldsymbol{L}|}{|\boldsymbol{\Sigma}|} - \frac{D}{2} + \frac{1}{2} \int p(\boldsymbol{x}) (\boldsymbol{x} - \boldsymbol{m})^T \boldsymbol{L}^{-1} (\boldsymbol{x} - \boldsymbol{m}) dx =
\end{aligned}
$$

Let's focus on solving the last term:

$$\begin{aligned}\int p(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{m})^T\boldsymbol{L}^{-1}(\boldsymbol{x}-\boldsymbol{m})dx &= \int p(\boldsymbol{x})\boldsymbol{x}^T\boldsymbol{L}^{-1}\boldsymbol{x}dx - 2\int p(\boldsymbol{x})\boldsymbol{x}^T\boldsymbol{L}^{-1}\boldsymbol{m}dx - \int p(\boldsymbol{x})\boldsymbol{m}^T\boldsymbol{L}^{-1}\boldsymbol{m}dx = \\ &= +Tr\Big[\boldsymbol{L}^{-1}\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^T]\Big] - 2Tr\Big[\boldsymbol{L}^{-1}\boldsymbol{m}\mathbb{E}[\boldsymbol{x}]\Big] + \boldsymbol{m}^T\boldsymbol{L}^{-1}\boldsymbol{m} = \\ &= +Tr\Big[\boldsymbol{L}^{-1}(\boldsymbol{\mu}\boldsymbol{\mu}+\boldsymbol{\Sigma})\Big] - 2Tr\Big[\boldsymbol{L}^{-1}\boldsymbol{m}\boldsymbol{\mu}^T\Big] + \boldsymbol{m}^T\boldsymbol{L}^{-1}\boldsymbol{m} = \\ &= \boldsymbol{\mu}^T\boldsymbol{L}^{-1}\boldsymbol{\mu} + Tr[\boldsymbol{L}^{-1}\boldsymbol{\Sigma}] - 2\boldsymbol{\mu}^T\boldsymbol{L}^{-1}\boldsymbol{m} + \boldsymbol{m}^T\boldsymbol{L}^{-1}\boldsymbol{m} = \\ &= \boldsymbol{\mu}^T\boldsymbol{L}^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\boldsymbol{L}^{-1}\boldsymbol{m} + \boldsymbol{m}^T\boldsymbol{L}^{-1}\boldsymbol{m} + Tr[\boldsymbol{L}^{-1}\boldsymbol{\Sigma}] = \\ &= (\boldsymbol{\mu}-\boldsymbol{m})^T\boldsymbol{L}^{-1}(\boldsymbol{\mu}-\boldsymbol{m}) + Tr[\boldsymbol{L}^{-1}\boldsymbol{\Sigma}]\end{aligned}$$

Returning to the computation of the KL-divergence:

$$\begin{aligned}\mathrm{KL}\big(p||q\big) &= -\int p(\boldsymbol{x})\log\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}dx = \\ &= \frac{1}{2}\log\frac{|\boldsymbol{L}|}{|\boldsymbol{\Sigma}|} - \frac{D}{2} + \frac{1}{2}\int p(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{m})^T\boldsymbol{L}^{-1}(\boldsymbol{x}-\boldsymbol{m})dx = \\ &= \frac{1}{2}\Big(\log\frac{|\boldsymbol{L}|}{|\boldsymbol{\Sigma}|} - D + (\boldsymbol{\mu}-\boldsymbol{m})^T\boldsymbol{L}^{-1}(\boldsymbol{\mu}-\boldsymbol{m}) + Tr[\boldsymbol{L}^{-1}\boldsymbol{\Sigma}]\Big)\end{aligned}$$

3.2 Question 2

Reminding that $\int p(\boldsymbol{x}) = \int \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = 1$:

$$\begin{aligned}H(p) &= -\int p(\boldsymbol{x})\log p(\boldsymbol{x})dx = \\ &= \frac{D}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}\int p(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})dx = \\ &= \frac{D}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}\int p(\boldsymbol{x})Tr\Big(\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\Big)dx = \\ &= \frac{D}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}Tr\Big(\boldsymbol{\Sigma}^{-1}\int p(\boldsymbol{x})(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T dx\Big) = \\ &= \frac{D}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}Tr\Big(\boldsymbol{\Sigma}^{-1}Cov(\boldsymbol{x},\boldsymbol{x})\Big) = \\ &= \frac{D}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}Tr\Big(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\Big) = \\ &= \frac{D}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{1}{2}Tr\Big(\boldsymbol{I}\Big) = \\ &= \frac{D}{2}\log 2\pi + \frac{1}{2}\log|\boldsymbol{\Sigma}| + \frac{D}{2} = \\ &= \frac{1}{2}\log\Big((2e\pi)^D|\boldsymbol{\Sigma}|\Big)\end{aligned}$$